EPIDP8451 Final Project - An application of machine learning to identify potential risk factors for early childhood asthma

Fang Liu (fl2569), Jing Wen (jw4061)

4/13/2022

Table of Contents

# Reseach Background

Option 3: group assignment Dataset: Exposome

## Research Question

What are the risk factors for predicting asthma diagnosis in children of early school age?

## Research Rationale

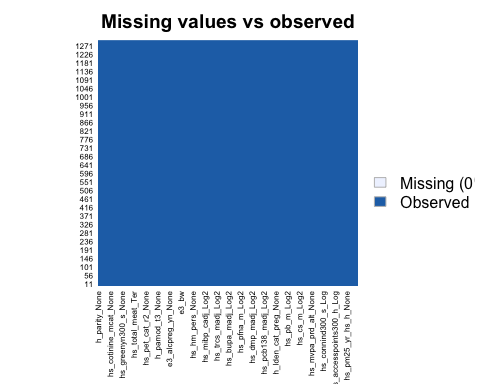
Asthma is the most common chronic airway disease of childhood. The etiology of asthma is not fully understood, but family history of the disease, respiratory infections, as well as environmental factors (e.g. airborne pollutants) may contribute to disease development. This study aims to identify modifiable risk factors of asthma for predicting asthma during early school age.

# Part 1 - Data preparation

## Data Cleaning & Preprocessing

In this step, we performed data cleaning for final dataset which included: 1) load data, 2) change categorical variables to factors, 3) strip ID variable, 4) check for missing values, 5) check if there is imbalance of outcome distribution, and 6) address collinearity by removing features with >0.8 correlation before modeling.

#Load data  
load("./exposome.RData")  
data <- merge(exposome,phenotype,by="ID") %>% merge(covariates, by="ID")  
  
#Strip off ID Variable  
data$ID<-NULL  
  
#make variable as factor  
data$hs\_asthma <- as.factor(data$hs\_asthma) #241 features  
  
# Finding correlated predictors  
data\_numeric <- data %>% select(where(is.numeric)) #180 features  
data\_categorical <- data %>% select(!where(is.numeric)) #61 features  
  
correlations <- cor(data\_numeric, use="complete.obs")  
high.correlations <- findCorrelation(correlations, cutoff=0.8)   
  
#findCorrelation() searches through a correlation matrix and returns a vector of integers corresponding to COLUMNS to remove to reduce pair-wise correlations  
  
# Remove highly correlated features  
data\_numeric\_low\_corr <- data\_numeric[,-high.correlations] #159 variables (continuous)  
  
#Combine the low-correlated numerical variables & categorical variables  
final\_data = bind\_cols(data\_numeric\_low\_corr, data\_categorical)   
#1301 observations x 220 variables columns  
  
#check missing data  
missmap(final\_data, main = "Missing values vs observed", x.cex = 0.5, y.cex = 0.5)



#check if data is balanced  
summary(final\_data$hs\_asthma) #very unbalanced 1159 vs. 142

## 0 1   
## 1159 142

**Conclusion:** In the original dataset, there are 241 features (180 numerical variables + 61 categorical variables); We removed 21 highly correlated numeric variables (180-21=159), so we are left with **220** variables for feature selection for the next step.

## Data Partition

In this step, models would be trained on 70% of patient data (the training set) and tested on the remaining 30% (the test set) as an internal validation of the model.

set.seed(100)  
train\_index <- createDataPartition(y=final\_data$hs\_asthma, p=0.7, list=FALSE)  
train\_data <- final\_data[train\_index,] #912  
test\_data <- final\_data[-train\_index,] #389

# Part 2 - Variable Selection

## Variable Selection using random forest

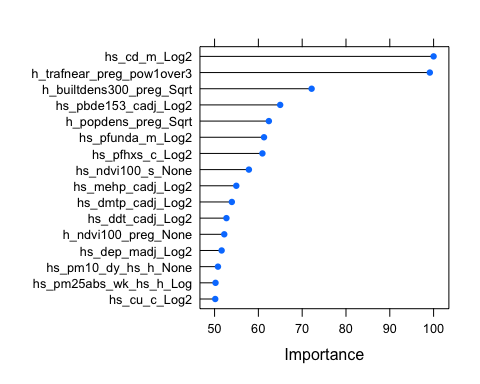
set.seed(100)  
  
#hyper parameter tuning: mtry  
mtry\_val2 <- c(ncol(train\_data)-1, sqrt(ncol(train\_data)-1), 0.5\*ncol(train\_data)-1)  
mtry\_grid <- expand.grid(.mtry=mtry\_val2)   
  
#5-fold cross-validation with upsampling  
rf\_asthma <- train(hs\_asthma ~ ., data=train\_data, method="rf", trControl=trainControl("cv", number=5, sampling="up"), metric="Accuracy", tuneGrid=mtry\_grid, ntree=100)  
  
rf\_asthma$bestTune

## mtry  
## 1 14.79865

rf\_asthma$results

## mtry Accuracy Kappa AccuracySD KappaSD  
## 1 14.79865 0.8892572 -0.002103602 0.002257404 0.004703798  
## 2 109.00000 0.8837687 0.015334993 0.006086021 0.031890455  
## 3 219.00000 0.8739086 0.011565498 0.012211379 0.034041651

#find which variables are considered "important"  
varlist = varImp(rf\_asthma)[["importance"]]  
important\_vars = varlist %>% filter(Overall > 50) %>% arrange(-Overall)  
  
#plot the important variables  
plot(varImp(rf\_asthma), top=16)



#paste(rownames(important\_vars), collapse = " + ")

**Conclusion:** Feature selection was done by a first round of 5-fold cross validated Random Forest model on the full predictor set after removing highly-correlated features (220 features). We used the top 16 features from this model’s feature importance (i.e., individual features with 50% importance or higher were selected). A complete list of predictor variables is as followed: hs\_cd\_m\_Log2, h\_trafnear\_preg\_pow1over3, h\_builtdens300\_preg\_Sqrt, hs\_pbde153\_cadj\_Log2, h\_popdens\_preg\_Sqrt, hs\_pfunda\_m\_Log2, hs\_pfhxs\_c\_Log2, hs\_ndvi100\_s\_None, hs\_mehp\_cadj\_Log2, hs\_dmtp\_cadj\_Log2, hs\_ddt\_cadj\_Log2, h\_ndvi100\_preg\_None, hs\_dep\_madj\_Log2, hs\_pm10\_dy\_hs\_h\_None, hs\_pm25abs\_wk\_hs\_h\_Log, hs\_cu\_c\_Log2.

# Part 3 - Construct prediction models

## Model Building & Evaluation

In this step, several machine-learning methods were applied and assessed to predict asthma diagnosis. These methods include traditional logistic regression, Logistic Regression with regularization (Elastic Net), Support Vector Machine (SVM) and ensemble Bagging. To reduce the problem of imbalanced outcome distribution, imbalance corrections were used with Random upsampling which showed the better performance than down-sampling. In addition, to improve the robustness and avoid overfitting (lack of generalization), we used 10-fold cross-validation with group split for the model selection process on the training data.

### Reference Model: Traditional Logistic regression

Traditional logistic regression was used as reference. Since we were concerned about the over-fitting problem from logistic model, the machine learning models below were presented to avoid over-fitting and prediction.

set.seed(100)  
  
logit\_asthma <- train(hs\_asthma ~ hs\_cd\_m\_Log2 + h\_trafnear\_preg\_pow1over3 + h\_builtdens300\_preg\_Sqrt + hs\_pbde153\_cadj\_Log2 + h\_popdens\_preg\_Sqrt + hs\_pfunda\_m\_Log2 + hs\_pfhxs\_c\_Log2 + hs\_ndvi100\_s\_None + hs\_mehp\_cadj\_Log2 + hs\_dmtp\_cadj\_Log2 + hs\_ddt\_cadj\_Log2 + h\_ndvi100\_preg\_None + hs\_dep\_madj\_Log2 + hs\_pm10\_dy\_hs\_h\_None + hs\_pm25abs\_wk\_hs\_h\_Log + hs\_cu\_c\_Log2, data = train\_data, method="glm", family = "binomial", trControl=trainControl(method="cv", number = 10))  
  
logit\_asthma$results

## parameter Accuracy Kappa AccuracySD KappaSD  
## 1 none 0.8892499 -0.002038736 0.003593446 0.006447049

### Model 1: Elastic Net (regularized regression)

Elastic net (EN) was used because it is a hybrid of ridge and lasso regularization. We used regularized regression to avoid the over-fitting problem from traditional regression. It’s also useful for identifying important features. The tuning hyperparameter alpha and lambda in our EN model is **0.8** and **0.00115**, respectively.

set.seed(100)  
  
en\_asthma <- train(hs\_asthma ~ hs\_cd\_m\_Log2 + h\_trafnear\_preg\_pow1over3 + h\_builtdens300\_preg\_Sqrt + hs\_pbde153\_cadj\_Log2 + h\_popdens\_preg\_Sqrt + hs\_pfunda\_m\_Log2 + hs\_pfhxs\_c\_Log2 + hs\_ndvi100\_s\_None + hs\_mehp\_cadj\_Log2 + hs\_dmtp\_cadj\_Log2 + hs\_ddt\_cadj\_Log2 + h\_ndvi100\_preg\_None + hs\_dep\_madj\_Log2 + hs\_pm10\_dy\_hs\_h\_None + hs\_pm25abs\_wk\_hs\_h\_Log + hs\_cu\_c\_Log2,data=train\_data, method="glmnet", family="binomial", trControl = trainControl("cv", number = 10, sampling= "up"), tuneLength=10)  
  
en\_asthma$bestTune

## alpha lambda  
## 52 0.8 0.001150063

### Model 2: Support Vector Machine

Support Vector Machine (SVM) was used because it is capable of performing classification, regression and outlier detection. C is the tuning hyperparameter that controls how much margin can be violated (i.e. how much classifications). Tuning parameter C in our SVM model is **1.333667**.

set.seed(100)  
  
#train model   
svm\_asthma <- train(hs\_asthma ~ hs\_cd\_m\_Log2 + h\_trafnear\_preg\_pow1over3 + h\_builtdens300\_preg\_Sqrt + hs\_pbde153\_cadj\_Log2 + h\_popdens\_preg\_Sqrt + hs\_pfunda\_m\_Log2 + hs\_pfhxs\_c\_Log2 + hs\_ndvi100\_s\_None + hs\_mehp\_cadj\_Log2 + hs\_dmtp\_cadj\_Log2 + hs\_ddt\_cadj\_Log2 + h\_ndvi100\_preg\_None + hs\_dep\_madj\_Log2 + hs\_pm10\_dy\_hs\_h\_None + hs\_pm25abs\_wk\_hs\_h\_Log + hs\_cu\_c\_Log2, data = train\_data, method="svmLinear", trControl=trainControl(method="cv", number = 10, sampling="up"), preProcess = c("center","scale"), tuneGrid=expand.grid(C=seq(0.001,2,length=10)))  
  
svm\_asthma$bestTune

## C  
## 7 1.333667

### Model 3 - Bagging (ensemble)

Bagging was used because it’s the bootstrap aggregation that we can get average results across bootstrapped samples of training data. Compared to the random forest, all features are eligible for selection in bagging, which allows all features to contribute to prediction. Tuning parameter mtry in our bagging model is **14.79865**.

set.seed(100)  
  
#Note: in bagging, ALL predictor features are eligible for selection at each node   
#parameter tuning  
mtry\_val1 <- expand.grid(.mtry = c(ncol(train\_data)-1, sqrt(ncol(train\_data)-1), 0.5\*ncol(train\_data)-1))  
  
bag\_asthma<-train(hs\_asthma ~ hs\_cd\_m\_Log2 + h\_trafnear\_preg\_pow1over3 + h\_builtdens300\_preg\_Sqrt + hs\_pbde153\_cadj\_Log2 + h\_popdens\_preg\_Sqrt + hs\_pfunda\_m\_Log2 + hs\_pfhxs\_c\_Log2 + hs\_ndvi100\_s\_None + hs\_mehp\_cadj\_Log2 + hs\_dmtp\_cadj\_Log2 + hs\_ddt\_cadj\_Log2 + h\_ndvi100\_preg\_None + hs\_dep\_madj\_Log2 + hs\_pm10\_dy\_hs\_h\_None + hs\_pm25abs\_wk\_hs\_h\_Log + hs\_cu\_c\_Log2, data=train\_data, method="rf", metric="Accuracy", trControl = trainControl("cv", number = 10, sampling="up"), tuneGrid=mtry\_val1, ntree=100)   
  
bag\_asthma$bestTune

## mtry  
## 1 14.79865

## Model comparison

In this step, accuracy was used for model comparison because we already addressed the issue of imbalanced data using upsampling. The model with the highest accuracy value was selected as the final model and the testing data will be used to evaluate the model performance

#compare model performance using accuracy   
  
confusionMatrix(logit\_asthma) #Reference model: 0.8893

## Cross-Validated (10 fold) Confusion Matrix   
##   
## (entries are percentual average cell counts across resamples)  
##   
## Reference  
## Prediction 0 1  
## 0 88.9 11.0  
## 1 0.1 0.0  
##   
## Accuracy (average) : 0.8893

confusionMatrix(en\_asthma) #0.6338

## Cross-Validated (10 fold) Confusion Matrix   
##   
## (entries are percentual average cell counts across resamples)  
##   
## Reference  
## Prediction 0 1  
## 0 56.1 3.7  
## 1 32.9 7.2  
##   
## Accuracy (average) : 0.6338

confusionMatrix(svm\_asthma) #0.5833

## Cross-Validated (10 fold) Confusion Matrix   
##   
## (entries are percentual average cell counts across resamples)  
##   
## Reference  
## Prediction 0 1  
## 0 51.3 3.9  
## 1 37.7 7.0  
##   
## Accuracy (average) : 0.5833

confusionMatrix(bag\_asthma) #0.8761

## Cross-Validated (10 fold) Confusion Matrix   
##   
## (entries are percentual average cell counts across resamples)  
##   
## Reference  
## Prediction 0 1  
## 0 86.2 9.5  
## 1 2.9 1.4  
##   
## Accuracy (average) : 0.8761

**Conclusion:**  The accuracy measures for each of the models above are as follows: elastic net (**0.6338**), support vector machine (**0.5833**), bagging (**0.8761**); note that the traditional logistic regression model is included as reference. Because the ensemble bagging algorithm has the highest accuracy out of the three machine learning models, the bagging model is chosen as our final model and evaluated using the testing data.

# Part 4 - Model Evaluation (Bagging)

## Evaluate model performance using testing data

asthma\_pred = predict(bag\_asthma, test\_data)  
asthma\_pred\_prob = predict(bag\_asthma, test\_data, type = "prob")   
  
#Accuracy  
confusionMatrix(asthma\_pred, test\_data$hs\_asthma, positive = "1")

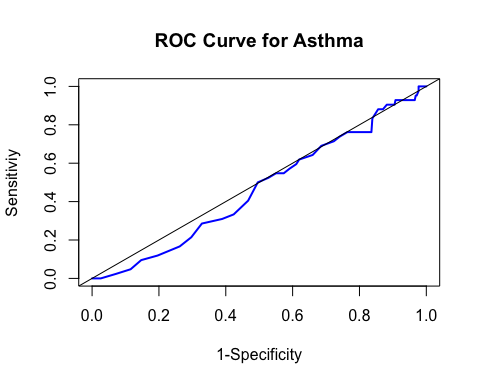
## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 336 40  
## 1 11 2  
##   
## Accuracy : 0.8689   
## 95% CI : (0.8312, 0.9008)  
## No Information Rate : 0.892   
## P-Value [Acc > NIR] : 0.9364   
##   
## Kappa : 0.0229   
##   
## Mcnemar's Test P-Value : 8.826e-05   
##   
## Sensitivity : 0.047619   
## Specificity : 0.968300   
## Pos Pred Value : 0.153846   
## Neg Pred Value : 0.893617   
## Prevalence : 0.107969   
## Detection Rate : 0.005141   
## Detection Prevalence : 0.033419   
## Balanced Accuracy : 0.507959   
##   
## 'Positive' Class : 1   
##

#AUC   
bag\_auc = roc(response=test\_data$hs\_asthma, predictor=asthma\_pred\_prob[,2])

## Setting levels: control = 0, case = 1

## Setting direction: controls > cases

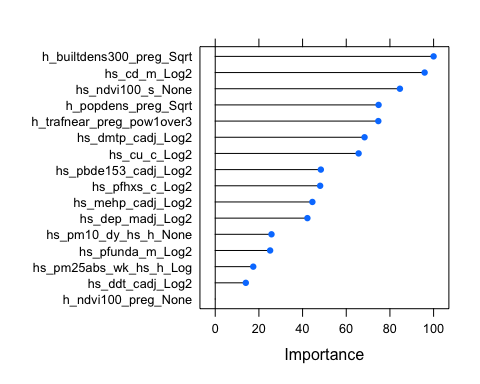
plot(1-bag\_auc$specificities, bag\_auc$sensitivities, type="l", ylab="Sensitiviy",xlab="1-Specificity",col="blue",lwd=2, main = "ROC Curve for Asthma")  
abline(a=0,b=1)



bag\_auc$auc

## Area under the curve: 0.4646

#Variable importance  
plot(varImp(bag\_asthma))



**Conclusion:**  Recall from pt 3) that ensemble bagging was selected as the final model because it had the highest accuracy out of the three machine learning models that we trained on the training set. Accuracy was **0.8761** for the training set and **0.8689** (95% CI: 0.8312, 0.9008) for the testing set. Area under the curve (AUC) was **0.4645945** with sensitivity and specificity of **0.047619** and **0.9683**, respectively. Model precision, also known as the positive predictive value(PPV) is **0.153846** and model recall is **0.047619**.

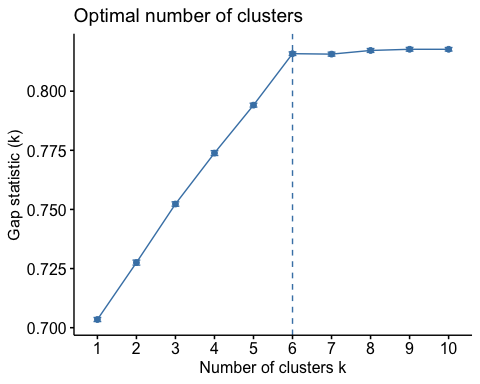
According to our variable importance results above, the 5 most important factors for predicting asthma diagnosis are: **building density during pregnancy, Cadmium (Cd) levels in mother during pregnancy, Average of NDVI values within a buffer of 100m at school, population density at pregnancy period, and Traffic density on nearest road during pregnancy.** Other important predictors include Copper (Cu) levels in child, exposure to Polybrominated diphenyl ether-153 (PBDE-153), Diethyl phosphate (DEP) in mother, and other environmental risk factors. Note that variable importance only tells us the relative importance for each variable and not the direction of the effect on our outcome of interest; in other words, we don’t know the direction of the association between the variables (e.g., is risk for asthma lower or higher with higher building density during pregnancy?).

# Part 5 - Unsupervised (k-means)

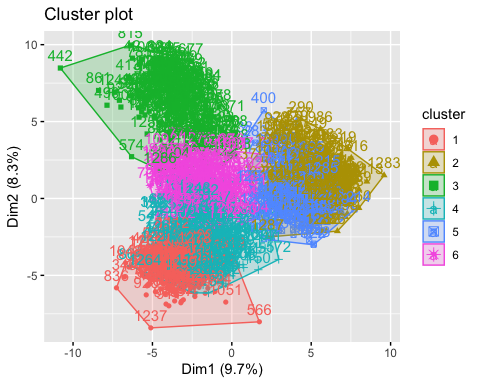
set.seed(100)  
  
# Center and scale data  
set.up.preprocess<-preProcess(data\_numeric\_low\_corr, method=c("center", "scale"))  
transformed.vals<-predict(set.up.preprocess, data\_numeric\_low\_corr)  
  
# Conduct a gap statistic analysis to determine optimal number of clusters   
gap\_stat<-clusGap(transformed.vals, FUN=kmeans, nstart=10, K.max=10, B=10, iter.max=50)   
print(gap\_stat, method="firstmax")

## Clustering Gap statistic ["clusGap"] from call:  
## clusGap(x = transformed.vals, FUNcluster = kmeans, K.max = 10, B = 10, nstart = 10, iter.max = 50)  
## B=10 simulated reference sets, k = 1..10; spaceH0="scaledPCA"  
## --> Number of clusters (method 'firstmax'): 6  
## logW E.logW gap SE.sim  
## [1,] 8.656995 9.360454 0.7034591 0.0008487758  
## [2,] 8.611550 9.339070 0.7275207 0.0010586672  
## [3,] 8.575648 9.327937 0.7522888 0.0009232700  
## [4,] 8.546067 9.319894 0.7738271 0.0010156081  
## [5,] 8.520129 9.314215 0.7940859 0.0008222115  
## [6,] 8.493462 9.309313 0.8158509 0.0007556343  
## [7,] 8.489414 9.305069 0.8156549 0.0008328795  
## [8,] 8.484022 9.301227 0.8172049 0.0008100967  
## [9,] 8.480133 9.297837 0.8177038 0.0007840171  
## [10,] 8.476944 9.294644 0.8176997 0.0007455337

fviz\_gap\_stat(gap\_stat)



#perform k-means & visualize the clusters  
clusters <- kmeans(transformed.vals, 6, nstart=25)  
fviz\_cluster(clusters, data=transformed.vals)



#Show the mean value of features within each cluster  
clus\_centers = clusters$centers  
  
as.data.frame(clus\_centers) %>%   
 select(h\_builtdens300\_preg\_Sqrt, hs\_cd\_m\_Log2, hs\_ndvi100\_s\_None, h\_popdens\_preg\_Sqrt, h\_trafnear\_preg\_pow1over3)

## h\_builtdens300\_preg\_Sqrt hs\_cd\_m\_Log2 hs\_ndvi100\_s\_None h\_popdens\_preg\_Sqrt  
## 1 -0.530362233 -0.12563998 0.8286938 -0.06482214  
## 2 0.004060833 -0.36452684 0.4843580 0.28741295  
## 3 1.093244477 -0.01552362 -1.2734890 1.02593147  
## 4 0.231174928 0.75304192 -0.7604185 -0.27699064  
## 5 -0.810369299 -0.10103865 0.4934964 -0.98925162  
## 6 -0.076536576 -0.00860776 0.1414524 -0.17288490  
## h\_trafnear\_preg\_pow1over3  
## 1 0.8508459  
## 2 -0.4982071  
## 3 1.0940114  
## 4 -0.6325920  
## 5 -0.1439868  
## 6 -0.5670504

#combine original data and the clustering results   
data\_with\_cluster <- cbind(data\_numeric\_low\_corr,cluster=clusters$cluster)

**Conclusion:**  An unsupervised K-means algorithm is used to group the 1,301 mother-child pairs based on their exposure characteristics during and post pregnancy (NOTE: only low-correlated numerical variables are included because k-means utilizes distance measures so categorical variables are excluded). The optimal # of clusters is chosen using the Gap Statistic, which is **6**. In brief, the basic idea of the Gap Statistic is to choose the number of K, where the biggest jump in within-cluster distance occurred. The cluster centers can be used to check if the observations in a cluster is generally above or below average for certain exposures. To illustrate, the observations in cluster #1 are below average in building density, mother cadium exposure, and population density, but is higher in NDVI levels are traffic density; those in cluster #4 are above average in building density and mother cadium levels during pregancy; but have lower levels of exposure to ndvi, lower population and traffic density. There is not really a true way to evaluate the “accuracy” of the clustering results since there is no outcome; nonetheless, findings from these preliminary clustering steps could be used to inform future research questions.

# Discussion

In this observational cohort study, by using several machine learning algorithms, we identified prediction factors for asthma diagnosis in children, some previously known and others novel. The top factors included building density during pregnancy, Cadmium (Cd) levels in mother during pregnancy, Average of NDVI values within a buffer of 100m at school, population density at pregnancy period, and Traffic density on nearest road during pregnancy. Model evaluation on test data yielded an accuracy of 0.8689 but an AUROC = 0.4646, indicating that the the variables we selected in our model alone may not be sufficient for asthma prediction. Nonetheless, our results indicated that exposure to environmental toxicants, especially in more densely-populated areas, and the nearby environment could be tied to the development of childhood asthma. To improve the health outcomes of adolescents, countries need heightened efforts to tighten the regulation of environmental pollution policies and improve the air quality for future generations.

The strengths of our study include a large sample size and multiple machine learning methods were fitted and compared with traditional logistic regression approach. The methods were compared based on accuracy and the method with the highest score was selected as final model. Additionally, feature selection was performed before modeling which contributed to a better prediction for the model building part. Limitations of the study include a small testing set and a lack of external data set for validation. This could reduce the generalizability of our model and might only apply to similar healthcare settings in other countries within Europe. As we are unaware of the data collection process, ethical issues may arise if certain population is underrepresented or excluded from the study. In addition, even though our final bagging model had a decent accuracy, the model precision and recall is very low, so is the AUC. Thus, this tool needs to be farther developed and evaluated. Additional data on environmental triggers such as weather, pollen count, and air quality might be necessary to improve performance of the short-term predictive model to develop a more useful tool.